

Fair Play or False Start?

What we are learning about large language model (LLM)-generated survey and assessment response data

Margarita Olivera-Aguilar, AIR[®]; Sam Rikoon, AIR[®]



Why This Work Matters

- High-quality surveys and assessments are essential for understanding what students know, can do, and experience, but testing their quality is expensive and time-consuming.
- By asking LLMs to mimic human responses when responding to surveys and assessments—that is, producing LLM-generated data—researchers could make early testing faster and less costly, especially before full pilot samples are available.
- A key question is whether LLM-generated data behave closely enough to real student responses, whereby such data could be used to test the measurement properties of surveys and assessments.

Bottom Line:

LLM-generated data may be a useful tool in early validation studies to pilot and refine new instruments more quickly and efficiently, but they should not fully replace human pilot studies.



What the Project Did

- Compared LLM-generated responses with real student data from four education measures: two surveys and two academic assessments.
- Tested more than 29 ways of prompting ChatGPT, with prompts focusing on student traits, skill levels, response styles, guessing, and expected performance. Generated 500 LLM-generated datasets for selected prompts to avoid findings dependent upon a single AI-driven sample.
- Checked (1) whether LLM-generated data reproduced broad response patterns from human data and (2) the measurement properties of the instruments being examined.



Key Learnings So Far

- *Prompting approach matters.* Only three prompts produced realistic survey item response patterns (e.g., the LLM used all available response options).
- *Surveys and tests should use different approaches.* Survey prompts worked better when they asked ChatGPT to respond with reference to a stable underlying trait. In one of the assessments, prompts worked better when they included expected skill levels, item difficulty, and student-like test-taking strategies.
- *LLM-generated data captured big-picture patterns.* Some LLM-generated datasets recovered the general latent structure observed in real student data and were useful for identifying items that required further review.
- *Fine-grained details remain difficult.* LLM-generated data were less reliable for reproducing the measurement properties of the items, especially item difficulty, as observed in human data.



What's Coming Next

- Continue data generation and analysis with additional LLMs (i.e., Claude, Llama, and Gemini) and instruments.
- Examine whether LLM-generated data contain the same types of biases (e.g., across sample groups) that are observed in human data.
- Investigate whether combining human and LLM-generated data can strengthen pilot studies with small sample sizes of humans.



1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
+1.202.403.5000 | [AIR.ORG](https://www.air.org)

Established in 1946, the American Institutes for Research® (AIR®) is a nonpartisan, not-for-profit institution that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of education, health, and the workforce. With headquarters in Arlington, Virginia, AIR has offices across the U.S. and abroad. For more information, visit [AIR.ORG](https://www.air.org).